ORIGINAL ARTICLE

Inter-rater agreement in the assessment of exposure to carcinogens in the offshore petroleum industry

Kjersti Steinsvåg, Magne Bråtveit, Bente E Moen, Hans Kromhout

Occup Environ Med 2007;**64**:582–588. doi: 10.1136/oem.2006.030528

See end of article for authors' affiliations

Correspondence to:
Kjersti Steinsvåg, University of Bergen, Department of Public Health and Primary Health Care, Section for Occupational Medicine, Kalfarveien 31, N-5018 Bergen, Norway; kjersti.steinsvag@isf.uib.no

Accepted 14 December 2006
Published Online First 16 January 2007

Objectives: To evaluate the reliability of an expert team assessing exposure to carcinogens in the offshore petroleum industry and to study how the information provided influenced the agreement among raters.
Methods: Eight experts individually assessed the likelihood of exposure for combinations of 17 carcinogens, 27 job categories and four time periods (1970–1979, 1980–1989, 1990–1999 and 2000–2005). Each rater assessed 1836 combinations based on summary documents on carcinogenic agents, which included descriptions of sources of exposure and products, descriptions of work processes carried out within the different job categories, and monitoring data. Inter-rater agreement was calculated using Cohen's kappa index and single and average score intraclass correlation coefficients (ICC) ($ICC_{(2,1)}$ and $ICC_{(2,8)}$, respectively). Differences in inter-rater agreement for time periods, raters, International Agency for Research on Cancer groups and the amount of information provided were consequently studied.
Results: Overall, 18% of the combinations were denoted as possible exposure, and 14% scored probable exposure. Stratified by the 17 carcinogenic agents, the probable exposure prevalence ranged from 3.8% for refractory ceramic fibres to 30% for crude oil. Overall mean kappa was 0.42 ($ICC_{(2,1)}=0.62$ and $ICC_{(2,8)}=0.93$). Providing limited quantitative measurement data was associated with less agreement than for equally well described carcinogens without sampling data.
Conclusion: The overall κ and single-score ICC indicate that the raters agree on exposure estimates well above the chance level. The levels of inter-rater agreement were higher than in other comparable studies. The average score ICC indicates reliable mean estimates and implies that sufficient raters were involved. The raters seemed to have enough documentation on which to base their estimates, but provision of limited monitoring data leads to more incongruence among raters. Having real exposure data, with the inherent variability of such data, apparently makes estimating exposure in a rigid semiquantitative manner more difficult.

Quantitative measurements of personal exposure to contaminants through air monitoring, skin deposition or biomonitoring are considered to be the most valid way to assess occupational exposure in epidemiological studies. For many cohort studies and essentially all case–control studies, which are retrospective, collecting reliable and valid retrospective exposure data is a challenge. To compensate for this, several proxy measures of exposure have been used such as job-exposure matrices, self-reported exposure assessment or expert assessment. In some studies, elements of different methods are combined.[1]

The use of expert assessment has increased in recent decades. Occupational hygienists, chemists, engineers and other professionals are considered to understand occupational exposure better than workers do. However, the experts may not be familiar with the jobs and industries to be considered,[2] and their background may influence how they assess exposure.[3] Hawkins and Evans[4] showed that, without measurement data, experts tended to overestimate exposure. The reliability in agreement between the experts might be tested by different statistical methods. Kappa statistics[5 6] have been used for categorical measures of exposure,[7–11] whereas intraclass correlation coefficients (ICC)[12 13] are often presented for continuous estimates.[3 14–16]

In expert assessment of exposure, the information provided to the experts varies. Some studies assume that the experts' prior knowledge allows them to make reliable and valid estimates.[11] In other studies, the experts have either conducted walk-through surveys and/or interviewed key workers prior to estimating exposure[3 7 17] or have been provided with written quantitative or qualitative information[4 8 14 18] or both.[19]

In 1998, the Cancer Registry of Norway established a Norwegian offshore cohort comprising 27 986 former and present offshore workers who completed a questionnaire on job history, lifestyle and demographics.[20] The development of cancer in this cohort will be analysed over the coming years. Qualitative and quantitative data for known and suspected carcinogens were obtained through company visits comprising interviews of key workers and collection of written documentations, including sampling reports. This background information, published previously,[21 22] shows that the measured data are scarce. In particular, there is a lack of both quantitative and qualitative information for the 1970s and 1980s. Visits to all offshore platforms were not feasible, thus strengthening the need for close cooperation with experts in occupational hygiene in the offshore petroleum industry. An expert group of eight people was therefore established to assess exposure to 17 carcinogenic agents, mixtures and exposure circumstances for 27 defined job categories during four time periods: 1970–1979, 1980–1989, 1990–1999 and 2000–2005. Prior to this assessment the experts were provided with summary documents on 17 carcinogenic agent, which included descriptions of sources of exposure and products, descriptions of work processes carried out within the different job categories, and sampling data.

This study used three categories of likelihood of exposure (unlikely, possible and probable exposure). Cohen's kappa (κ) index[23] gives the exact proportion of agreement that cannot be expected by chance alone. ICC measures the proportion of the

Abbreviations: IARC, International Agency for Research on Cancer; ICC, intraclass coefficients total variability attributable to the object of measurement,[13] in this study the combinations of carcinogen, job category and time period. Expert-based exposure ratings are subjective estimates and may be biased.[14] The kappa statistic does not take into account the impact any bias may have between two raters. ICC, in contrast, reacts to both the degree of mean differences (bias) and the correlation between raters. ICC also gives the opportunity to evaluate the number of raters involved by estimating the average score ICC.[24] The present study assessed the inter-rater agreement of the experts using both kappa statistics and ICC. The impact of time periods, amount of information, International Agency for Research on Cancer (IARC) group and the composition of the expert group on the agreement was investigated.

This study evaluated the reliability of an expert team assessing exposure to carcinogens in the offshore petroleum industry in order to examine whether the information and the number of raters used was sufficient to give reliable estimates and studied how the characteristics of the information provided influenced the agreement among raters.

METHODS

Forms for individual expert assessment of exposure

During a 1-day session, eight experts individually assessed the likelihood of exposure (unlikely, possible or probable) to 17 carcinogens for 27 job categories[22] and four time periods (1970–1979, 1980–1989, 1990–1999 and 2000–2005), resulting in 1836 combinations per rater. Prior to the expert rating, three-dimensional forms were prepared with one cell for each combination of carcinogen, job category and time period.

Each member of the expert group scored the likelihood of exposure. The expert group comprised eight individuals: three occupational hygienists from the offshore industry, two occupational hygienists from consulting companies affiliated with the offshore industry and three university researchers with experience in offshore projects.

To familiarise the experts with the methods of the assessment, they were handed the structure of the blank forms with instructions and guidance for completion 14 days before the meeting. Exposure was divided into three probability categories.

(1) Unlikely: it is unlikely that workers were exposed.
(2) Possible: it is possible that workers were exposed, but the probability is low, or<50% of the workers were probably exposed.
(3) Probable: probably at least 50% of the workers were exposed.

It was stressed that the most important task was to identify job categories with ‘‘probable exposure’’ and to avoid unexposed groups being denoted as probably exposed. ‘‘Exposure’’ was defined as occurring when exposure levels for the respective job categories exceed the assumed background levels in the living quarters of offshore installations.

Background information on possible exposure had been obtained through company visits, including interviews of key personnel (n = 83) and collection of sampling reports (n = 118) and other relevant documentation (n = 329). The companies comprised eight oil companies, five drilling companies, three chemical suppliers, three maintenance, modification and operation contractors, and a catering service supplier. Monitoring reports had been found for seven agents (benzene, mineral oil mist and oil vapour, dust, asbestos fibres, refractory ceramic fibres, formaldehyde and tetrachloroethylene).[22] The personal exposure to oil mist and oil vapour during drilling (65 reports) has been published previously.[21] Descriptions of products containing carcinogens, exposure sources and processes carried out within the different job categories were extracted from the documentation collected and the interviews of key personnel, and summarised for each selected carcinogen. The content of the summary reports have been presented in previous studies.[21 22]

In the expert session, the method was first presented and discussed. The experts then completed their individual forms based both on the written background information for each carcinogen and their own competence and experience. For about every third agent, the expert group had a brief discussion to clear up any misunderstandings in how to complete the form.

Statistical analysis

Data was analysed using SPSS V.13.0 for Windows (SPSS Inc., Chicago, Illinois, USA).

Unlikely, possible and probable exposures were entered into an SPSS database as the numbers 0, 1 and 2, respectively. Agreement parameters were grouped by carcinogen, rater background, IARC group, amount of information and time period. The coding of amount of information into the following three categories was performed by a university researcher through a subjective evaluation of the quality and amount of the background information:[21 22] (1) well described carcinogens with sampling data, (2) well described carcinogens without sampling data, and (3) less described carcinogens.

To investigate inter-rater agreement, Cohen's kappa index[23] and ICC[12] were calculated. One kappa value for each pair of raters was calculated, totally 28 pairs for eight raters. The kappa statistics are presented as the mean and range of kappa for the relevant rater pairs. If one of a pair of raters had not scored in all possible levels (unlikely, possibly or probably exposure), the kappa value could not be estimated. For example, if one rater in a pair had only used the categories ‘‘unlikely’’ and ‘‘possible’’ exposure in his or her assessments and the other had assessed all three categories, the kappa could not be estimated. The number of missing pairs is specified in the relevant table.

Mean and range of the kappa values for the seven rater pairs corresponding to each rater was calculated in order to examine if there were apparent differences in agreement regarding years of experience of the rater.

One-way analysis of variance was performed on the kappa values to detect significant differences between the subgroups within the categories of time periods, raters, IARC groups and amount of information. To investigate significant differences, further Bonferroni post hoc tests were performed.

Case 2 ICCs[12] were calculated using a two-way random analysis of variance including a random effect for score per combination of each set of eight raters and a random rater effect for each of the eight raters. In this study the two ICC measures, single and average score ICCs ($ICC_{(2,1)}$ and $ICC_{(2,8)}$, respectively) are presented according to Shrout and Fleiss[12] and McGraw and Wong, [13] rather than that of Teshcke *et al*,[3] who uses the denotations ‘‘individual ICC’’ and ‘‘group ICC’’. The number ‘‘2’’ refers to case 2,[12] and 1 and 8 refers to the number of raters. The CIs of the single-score ICCs were investigated to detect significant differences between the subgroups within the categories of time period, raters, IARC groups and amount of information. Using ICC values involves assumptions of normally distributed residuals in the two-way analysis of variance.[6] ICCs were used in this study despite violations of these assumptions.

To examine if there had been any trends throughout the day in the agreement among the raters during the filling of forms, analysis of the $ICC_{(2,1)}$ and $ICC_{(2,8)}$ results were performed for groups of three carcinogens, corresponding to the order in which they were assessed.

Pearson correlation coefficients were estimated to examine the correlations between the inter-rater reliability measures and the prevalence of possible and probable exposure. According to Altman,[6] investigated variables should preferably be normally distributed. Shapiro–Wilk W tests were therefore performed to test for normality.

Comparing the subgroups within the categories carcinogen, time periods, raters, IARC groups and amount of information requires that the subgroups have homogeneous between-combination variance. The mean square of between-combination ($MS_{combination}$) and residual mean square ($MS_{residual}$) was obtained using two-way analysis of variance when estimating ICC. The numbers were used to calculate the between-combination variance ($\sigma^2_{combination}$):

$$\sigma^2_{combination} = \frac{MS_{combination} - MS_{residual}}{k}$$

where k = number of raters. *F* tests were conducted to test for significant differences in the between-combination variance. When these tests are conducted, it is assumed that the two populations under investigation are normally distributed.[6]

RESULTS

Prevalence

Table 1 shows the total number of cells the eight raters completed individually. Of the 1836 carcinogen, job category and time period combinations, the raters denoted 67% as unlikely, 18% as possible and 14% as probable exposure. The university expert with least experience in occupational hygiene (rater 3) rated the highest number of possible exposures (494) and the lowest number of probable exposures (198), which was quite similar to rater 7, an industry expert. The senior expert from the oil industry had the highest number of probable exposure cells (379) (table 1).

Table 2 shows the prevalence of possible and probable exposure for each carcinogen. Benzene and crude oil had highest prevalence of possible exposure (36.8% and 29.8%, respectively) and high prevalence of probable exposure (29.0% and 23.2%, respectively). Skin exposure to mineral and crude oil and exposure to benzene had highest prevalence of probable exposure (30.3%, 23.2% and 29.0%, respectively). Low prevalence of probable exposure (<10%) was rated for formaldehyde (5.7%), ionising radiation (8.8%), occupational exposure as a painter (4.6%), lead (7.9%), dichloromethane (3.7%) and refractory ceramic fibres (8.1%) (table 2).

The prevalence of possible exposure ranged from 18.8% to 19.3% for the three time periods before 2000, whereas 2000–2005 was slightly lower at 15.9% (table 3). For probable exposure, the experts assigned 16.0% of the cells for 1970–1979 and 17.0% for 1980–1989 versus 12.5% for 1990–1999 and 9.6% for 2000–2005.

Data comprise kappa (mean and range of n rater pairs), one-way analysis of variance of kappa values between category subgroups, single ($ICC_{(2,1)}$) and average ($ICC_{(2,8)}$) score with respective 95% CIs for ICC and between-combination variance ($\sigma^2_{combination}$) for groups of time periods, raters, IARC groups and amount of information provided.

The university and industry groups of experts had similar estimated prevalence of possible and probable exposure (table 3). IARC groups 1 and 2A had similar prevalence (around 19% possible and 13% probable exposed combinations). Although group 2B had low prevalence of 12% for possible and 5% for probable exposure, group 3 chemicals showed the highest prevalence (23% for possible and 25% for probable).

When grouped by amount of information, the experts had completed most combinations for ‘‘well described carcinogens with sampling data’’ (table 3).

Inter-rater reliability

The overall kappa was 0.42 (table 2). Silica had the lowest kappa ($\kappa = 0.27$), and the experts agreed most on nickel compounds ($\kappa = 0.52$).

The overall $ICC_{(2,1)}$ was 0.62, with formaldehyde scoring lowest (0.39) and nickel again highest (0.75) (table 2). However, the between-combination variance differed, ranging from 0.09 (refractory ceramic fibres) to 0.46 (mineral oil on the skin), indicating that ICC values might not be comparable.

The overall average score ICC of 0.93 indicates that the group mean exposure estimates were reliable and that the expert group was sufficiently large to make reliable mean estimates. The agreement measures given in table 3 for the three university raters ($ICC_{(2,3)} = 0.82$) and the five industry raters ($ICC_{(2,5)} = 0.89$), gives an indication of average score when reducing the number of raters.

There were no trends in the agreements ($ICC_{(2,1)}$ and $ICC_{(2,8)}$) when the carcinogens were grouped according to the order in which they were assessed.

The ICC per carcinogen followed approximately the tendency of the kappa value per agent. Pearson correlation between

Table 1 Information on the expert raters assessing exposure to 17 carcinogens for 27 job categories in four time periods in the offshore petroleum industry

Rater	Background	Years of occupational hygiene experience	Years of occupational hygiene experience offshore	No of unlikely cells (%)	No of possible cells (%)	No of probable cells (%)
1	University	25	3	1278 (69.6)	309 (16.8)	249 (13.6)
2	University	15	3	1304 (71.0)	276 (15.0)	256 (13.9)
3	University	3	3	1144 (62.3)	494 (26.9)	198 (10.8)
4	Industry*	19	3	1238 (67.4)	347 (18.9)	251 (13.7)
5	Industry*	23	3	1363 (74.2)	238 (13.0)	235 (12.8)
6	Industry†	15	15	1120 (61.0)	337 (18.4)	379 (20.6)
7	Industry†	13	8	1147 (62.5)	479 (26.1)	210 (11.4)
8	Industry†	6	4	1379 (75.1)	209 (11.4)	248 (13.5)
Mean		15	5	1247 (67.9)	336 (18.3)	253 (13.8)

Data included background, years of experience in occupational hygiene (totally and in offshore industry) and the number of cells scored ‘‘unlikely’’, ‘‘possibly’’ or ‘‘probably’’ exposed.
*Industry rater from contracting companies.
†Industry rater from oil companies.

Table 2 Carcinogens in the offshore petroleum industry evaluated by an expert group

Carcinogen	IARC groups	Amount of information	Exposure possible, (range) (%)	Exposure probable, (range) (%)	Kappa (range) (M*)	ICC$_{(2,1)}$ (95% CI)	ICC$_{(2,8)}$ (95% CI)	σ^2†
Asbestos	1	1	27.0 (15.7 to 37.0)	16.9 (3.70 to 45.4)	0.41 (0.09 to 0.67)	0.62 (0.54 to 0.70)	0.93 (0.90 to 0.95)	0.36
Benzene	1	1	36.8 (15.7 to 64.8)	29.0 (16.7 to 44.4)	0.35 (0.05 to 0.61)	0.57 (0.48 to 0.66)	0.91 (0.88 to 0.94)	0.37
Formaldehyde	1	1	14.2 (0 to 27.8)	5.68 (0 to 18.5)	0.30 (−0.09 to 0.86) (17)	0.39 (0.32 to 0.48)	0.84 (0.79 to 0.88)	0.12
Silica	1	1	27.6 (3.7 to 48.1)	14.9 (7.40 to 25.0)	0.27 (0.07 to 0.58)	0.43 (0.34 to 0.53)	0.86 (0.80 to 0.90)	0.24
Chromium (VI)	1	2	15.4 (3.7 to 45.4)	12.6 (8.3 to 57.4)	0.51 (0.16 to 0.76)	0.74 (0.67 to 0.80)	0.96 (0.94 to 0.97)	0.37
Ionising radiation	1	2	7.64 (0 to 14.8)	8.79 (3.7 to 18.5)	0.49 (0.26 to 0.88) (12)	0.64 (0.57 to 0.71)	0.93 (0.91 to 0.95)	0.24
Occupational exposure as a painter	1	2	6.61 (0 to 22.2)	4.64 (3.7 to 7.4)	0.50 (0.15 to 0.85) (7)	0.71 (0.65 to 0.77)	0.95 (0.94 to 0.96)	0.16
Nickel compounds	1	3	14.7 (0 to 43.5)	11.0 (8.3 to 13.9)	0.52 (0.20 to 0.87) (7)	0.75 (0.69 to 0.81)	0.96 (0.95 to 0.97)	0.34
Lead	2A	2	21.3 (8.3 to 57.4)	7.86 (3.7 to 12)	0.44 (0.15 to 0.69)	0.60 (0.51 to 0.69)	0.92 (0.89 to 0.95)	0.24
Chlorinated hydrocarbons	2A	3	19.8 (2.8 to 31.5)	23.4 (13.9 to 43.5)	0.40 (0.17 to 0.68)	0.58 (0.49 to 0.66)	0.92 (0.89 to 0.94)	0.41
Diesel engine exhaust	2A	3	16.3 (0 to 37)	9.05 (1.9 to 30.6)	0.32 (−0.02 to 0.68) (7)	0.43 (0.34 to 0.52)	0.86 (0.81 to 0.89)	0.18
Refractory ceramic fibres	2B	1	14.5 (0 to 44.4)	3.26 (1.9 to 5.6)	0.28 (0.07 to 0.61) (7)	0.41 (0.32 to 0.50)	0.84 (0.79 to 0.89)	0.09
Dichloromethane	2B	2	4.16 (0 to 8.3)	3.71 (0 to 11.1)	0.43 (0.19 to 0.65) (17)	0.56 (0.48 to 0.64)	0.91 (0.88 to 0.93)	0.10
Welding	2B	3	16.4 (2.8 to 40.7)	8.10 (2.80 to 14.4)	0.40 (0.09 to 0.78)	0.61 (0.52 to 0.69)	0.93 (0.90 to 0.95)	0.24
Mineral oil: inhalation	3	1	22.7 (5.6 to 48.1)	21.7 (14.8 to 27.8)	0.33 (0.14 to 0.69)	0.55 (0.46 to 0.63)	0.91 (0.87 to 0.93)	0.36
Mineral oil: skin	3	2	16.3 (0.9 to 37)	30.3 (18.5 to 57.4)	0.41 (0.17 to 0.71)	0.58 (0.48 to 0.67)	0.92 (0.88 to 0.94)	0.46
Crude oil: skin	3	3	29.8 (10.2 to 50.9)	23.2 (14.8 to 29.6)	0.37 (0.16 to 0.68)	0.55 (0.47 to 0.63)	0.91 (0.88 to 0.93)	0.36
Overall	—	—	—	—	0.42 (0.27 to 0.49)	0.62 (0.60 to 0.63)	0.93 (0.92 to 0.93)	0.32

Data included IARC groups, amount of information, percentage mean prevalence and range for possible and probable scored cells, mean and range of kappa for 28 rater pairs, single intraclass correlation coefficients (ICC$_{(2,1)}$), average score (ICC$_{(2,8)}$) and between-combinations variance ($\sigma^2_{combination}$) for 17 carcinogenic agents.
*M, number of missing rater pairs.
†$\sigma^2_{combination}$.

Table 3 Prevalence of expert-assessed possible and probable exposure to carcinogens in the offshore petroleum industry

Category	Exposure possible (%)	Exposure probable (%)	Kappa (range) (n)	Analysis of variance of kappa values	ICC$_{(2,1)}$ (95% CI)	ICC$_{(2,8)}$ (95% CI)	σ^2†
Time period							
1970–1979	19.3	16.0	0.41 (0.26–0.52) (28)	0.04	0.62 (0.58 to 0.65)	0.93 (0.92 to 0.94)	0.35
1980–1989	19.2	17.0	0.44 (0.30–0.54) (28)		0.64 (0.61 to 0.68)	0.93 (0.93 to 0.94)	0.38
1990–1999	18.8	12.5	0.40 (0.25–0.49) (28)		0.59 (0.55 to 0.63)	0.92 (0.91 to 0.93)	0.29
2000–2005	15.9	9.6	0.40 (0.26–0.49) (28)		0.59 (0.55 to 0.63)	0.92 (0.91 to 0.93)	0.25
Raters							
University	19.6	12.8	0.38 (0.27–0.47) (3)	0.27	0.60 (0.58 to 0.62)	0.82 (0.80 to 0.83)*	0.30
Industry	17.5	14.4	0.42 (0.38–0.49) (10)		0.62 (0.60 to 0.64)	0.89 (0.88 to 0.90)†	0.33
IARC groups							
1	18.7	12.9	0.44 (0.29–0.57) (28)	0.004	0.65 (0.62 to 0.67)	0.94 (0.93 to 0.94)	0.33
2A	19.1	13.5	0.37 (0.21–0.64) (28)		0.56 (0.51 to 0.61)	0.91 (0.89 to 0.92)	0.29
2B	11.7	5.0	0.35 (0.17–0.58) (28)		0.55 (0.50 to 0.59)	0.91 (0.89 to 0.92)	0.15
3	23.0	25.1	0.37 (0.23–0.56) (28)		0.56 (0.51 to 0.61)	0.91 (0.89 to 0.93)	0.39
Amount of information							
Well described carcinogens with sampling data	23.8	15.2	0.37 (0.19–0.49) (28)	<0.001	0.57 (0.54 to 0.60)	0.91 (0.90 to 0.92))	0.31
Well described carcinogens without sampling data	11.9	11.3	0.47 (0.33–0.60) (28)		0.67 (0.64 to 0.70)	0.94 (0.93 to 0.95)	0.30
Less described carcinogens	19.4	15.0	0.40 (0.29–0.57) (28)		0.61 (0.57 to 0.64)	0.92 (0.91 to 0.93)	0.33

*ICC$_{(2,3)}$: three university raters; †ICC$_{(2,5)}$: five industry raters; n, number of rater pairs.
†$\sigma^2_{combination}$.

kappa and $ICC_{(2,1)}$ per agent was 0.94 ($p<0.001$), whereas $ICC_{(2,8)}/ICC_{(2,1)}$ was 0.99 ($p<0.001$), and $ICC_{(2,8)}/\kappa$ was 0.92 ($p<0.001$). The measures of inter-rater agreement were not significantly correlated with the percentage mean prevalence of possible or probable exposure—that is, the prevalence of exposure did not seem to affect the agreement between the raters. Shapiro–Wilk W tests revealed that all the variables in the bivariate correlation estimations were normally distributed.

Factors influencing the agreement between raters

One-way analysis of variance (analysis of variance) showed significant differences in kappa values ($p = 0.04$) between the time periods (table 3). Post hoc tests indicated that the significant difference was between 1980–1989 and 1990–1999 ($p = 0.05$).

University experts and industry experts did not differ in agreement ($p = 0.27$) (table 3), and none of these groups differed significantly from the total kappa of 0.42.

IARC groups 2A, 2B and 3 had similar kappa values, whereas the agreement for IARC group 1 was slightly higher (table 3). Analysis of variance revealed significant differences between the groups, and Bonferroni post hoc tests indicated that the major differences were between IARC groups 1 and 2B ($p = 0.03$) and groups 1 and 3 ($p = 0.04$).

Analysis of variance was also performed for the three groups of carcinogens according to the amount of information available. The ''well described carcinogens with sampling data'' had the least agreement ($\kappa = 0.37$). Analysis of variance showed highly significant differences between the groups (table 3), and the Bonferroni post hoc test indicated a difference ($p<0.001$) between the two well described groups (with and without sampling data). The difference between the ''well described carcinogens without sampling data'' and the ''less described carcinogens'' groups was also significant ($p = 0.004$).

To evaluate differences between groups for $ICC_{(2,1)}$, the 95% CIs (table 3) were examined. For the time periods and raters, the CIs overlapped, indicating no outstanding subgroups. IARC group 1 had a 95% CI that was not similar to the other three groups, but *F* tests of the between-combination variance ($\sigma^2_{combination}$) differed significantly across the four IARC groups ($p<0.001$)—that is, the groups were not statistically comparable, making the conclusion of more agreement in IARC group 1 unfeasible. For the amount of information, agreement was higher in the ''well described carcinogens without sampling data'' subgroup than in the other two. There was least agreement for the ''well described carcinogens with sampling data'', and its 95% CI overlapped with the ''less described carcinogens'' group but not with the ''well described carcinogens without sampling data''. There was borderline overlap between the ''less described carcinogens'' group and the ''well described carcinogens without sampling data'' group. *F* tests revealed no significant differences between the variances for these groups except for the combination of ''less described carcinogens'' and ''well described carcinogens without sampling data'' ($p<0.05$).

No apparent differences were found in kappa agreement regarding years of experience of the raters.

DISCUSSION

Eight raters individually estimated exposure to 17 carcinogens in the offshore petroleum industry. For the 1836 exposure combinations assessed per rater, an overall kappa of 0.42 and single score ICC of 0.62 indicates that the agreement of raters in this study on exposure estimates was greater than chance. The lack of full agreement indicates that their subjective opinions influenced the decisions. The kappa values were in the upper range of comparable studies. In a study scoring the likelihood of exposure in three categories (0, unlikely; 1, possible; and 2, probable), van Tongeren *et al*[11] found an overall kappa between the raters of 0.36 for 0 versus 1 or 2, and 0.31 for 0 or 1 versus 2. The authors suggest that the poor agreement was due to lack of information on occupations and tasks. In a case–control study of brain tumour in which five experts assessed the presence or absence of exposure to 21 chemicals in 199 jobs, the kappa values for pairwise inter-rater agreement ranged from 0 to 0.6, with the median being $\kappa = 0.2$.[8]

The overall average score for the $ICC_{(2,8)}$ of 0.93 indicates reliable mean estimates of exposure and that the study included sufficient raters. Reduction from eight to five or three raters affected the average score ICC only marginally. An $ICC>0.81$ is defined as nearly perfect agreement.[5 24] The raters seem to have received enough information to give reliable average mean assessments. However, the industry raters represented the industrial sector under investigation, indicating that the assumption of independence between the raters might be questioned. A certain common understanding of exposure among occupational hygienists in this industry is expected as they often work on similar topics to comply with work environment regulations or, at times, to meet news headlines on chemical exposure. The occupational hygienists also arrange meetings to exchange and discuss mutual professional challenges, which might create a more homogeneous perception of exposure. The university experts did not differ from that of the industry experts. In accordance with Teschke's[1] recommendations, this study aimed at providing the experts with measurement data, information about the properties of the carcinogens, and detailed information about the worksite on which to base their likelihood estimates.

There is a rationale to conclude from the calculated kappa statistics and single-score ICC that providing limited quantitative data is associated with less agreement among raters than for equally well described carcinogens without sampling data. ICC estimates for different groups might not be comparable if the difference in between-subject variance is great. Analysis of the between-subject variance for the three categories of amount of information gave similar results, and therefore it is assumed that comparison is appropriate. Some studies have looked at changes in inter-rater agreement when providing their experts with cycles of increasing amount of information. In a study by de Cock *et al*,[15] information on pesticide exposure in fruit growing was provided to experts in three phases. The inter-rater agreement in ranking tasks with respect to exposure did not alter with increasing amount of information. Stewart *et al*[16] evaluated experts' assessments of formaldehyde exposure in manufacturing plants. Information on exposure was provided in six cycles of increasing amount of information, starting with job category and industry, and then adding dates, department title and plant reports. The mean difference between the hygienists' evaluations and a standard, more in-depth evaluation was slightly improved with increasing level of information (κ). When more quantitative information on captan exposure was given, the inter-rater agreement (κ) decreased.[15] However, according to Hawkins and Evans,[4] offering measurement data produces less biased expert estimates. They showed that, without measurement data, experts tend to overestimate exposure. When Post *et al*[19] gave measured data to occupational hygienists, their relative exposure ranking of jobs did not improve but their classification of jobs into quantitative exposure categories did, and agreement between the raters increased. Segnan *et al.*[18] compared assessments by experts, at different stages, based on occupational histories (median $ICC = 0.11$), industry-specific questionnaires (median $ICC = 0.21$), lists of products used (median $ICC = 0.65$), and where available, exposure measurement data (median ICC = 0.51). In general, increasing the information on monitoring data decreased agreement among the experts. The main reason for this is presumably the large inherent variability in individual measurement results.[25]

The exposure level to carcinogens in the offshore petroleum industry was generally low.[22] Other studies have discussed how the exposure level affects expert agreement. Macaluso *et al*[14] found low single-score ICC for exposure combinations were scored at low intensity despite high percentage concordance. These authors questioned whether expert-based exposure assessment is suitable for low exposure levels.

IARC group 1 carcinogens had significantly higher kappa values than the other IARC groups. To our knowledge , it has not previously been reported that experts are more likely to agree on established carcinogens (IARC group 1) than on less established carcinogens.

All experts were familiar with the production process and job categories. Expert-based exposure ratings are subjective estimates and may be biased.[14] Kappa does not take into account the impact any bias may have between two raters. ICC, in contrast, react to both the degree of mean differences (bias) and the correlation between raters—that is, the ICC decreases in response both to larger mean differences between raters (bias) and to lower correlation between the raters. A brief discussion was conducted for every third agent to calibrate the experts in order to reduce correlation and bias. Analysis did not reveal any time trends in agreement during the day of the exposure assessments. Benke *et al*[8] found a training effect in their study, as four of five experts identified more exposure in their first assessments than in repeated assessments.

Kappa does not take into account the degree of disagreement, thus, all disagreements are treated equally. When the categories are ordered, it may be preferable to give different weights to the disagreements according to the magnitude of the discrepancy.[6] Roberts and McNamee[26] focused on the limitation of the single summary-weighted kappa coefficients and suggested a symmetrical matrix of kappa-type coefficients instead. They proposed the method as being suitable for ordinal scale where there is no underlying continuum. Teschke *et al*[24] stated that ICC is comparable to kappa but is used when there is continuous data. According to Fleiss and Cohen,[27] ICC is the special case of weighted kappa when the categories are equally distributed along one dimension.

Although experts had been provided with information about the method used 14 days before the session, the summary documents contained much information to handle within a brief time frame. However, reading through the summary papers required relative little effort and provided all the raters with the same background information. The assessment scores of likelihood of exposure were accepted and were easy to apply. Nevertheless, the present study does not take into account the intensity, duration or frequency of exposure that other studies have performed.[3 14 16] The cost-effectiveness of this method in terms of low time consumption for the experts involved seems to be a prerequisite in the offshore petroleum industry.

CONCLUSION

The overall kappa and single-score ICC indicate that agreement on exposure estimates among the raters in this study was greater than chance. The levels of inter-rater agreement are higher than thosefound in comparable studies. The average score ICC indicates very reliable mean estimates and implies that sufficient raters were used. It seems that the raters were provided with enough documentation on which to base their estimates, but that providing limited monitoring data leads to more incongruence among raters. Having real exposure data, with the inherent variability of such data, apparently makes estimating exposure in a rigid semiquantitative manner more difficult.

ACKNOWLEDGEMENTS

Special thanks to the offshore petroleum industry occupational hygienists involved in the expert group: Bjørg Eli Hollund (X-lab, Bergen), Kristin Svendsen (Norwegian University of Science and Technology, Trondheim), Jakob Nærheim and Liv-Torill Austgulen (Statoil ASA) and Inger Margrethe Haaland (Norsk Hydro ASA). We are grateful for the support on statistical methods provided by Rolf Moe-Nilssen and Valborg Baste, University of Bergen.

Authors' affiliations

Kjersti Steinsvåg, Magne Bråtveit, Bente E Moen, University of Bergen, Department of Public Health and Primary Health Care, Section for Occupational Medicine, Bergen, Norway
Hans Kromhout, Utrecht University, Institute for Risk Assessment Sciences, Utrecht, the Netherlands

The Norwegian Oil Industry Association funded this study.

Competing interests: none declared

REFERENCES

1 **Teschke K**. Exposure surrogates: job-exposure matrices, self-reports, and expert evaluations. In Nieuwenhuijsen MJ, ed. *Exposure assessment in occupational and environmental epidemiology*. Oxford: Oxford University Press, 2003:118–32.
2 **Teschke K**, Olshan AF, Daniels JL, *et al.* Occupational exposure assessment in case-control studies: opportunities for improvement. *Occup Environ Med* 2002;**59**:575–94.
3 **Teschke K**, Hertzman C, Dimich-Ward H, *et al.* A comparison of exposure estimates by worker raters and industrial hygienists. *Scand J Work Environ Health*, 1989;**5**:424–9.
4 **Hawkins NC**, Evans JS. Subjective estimation of toluene exposures: a calibration study of industrial hygienists. *Appl Ind Hyg* 1989;**4**:61–8.
5 **Landis JR**, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
6 **Altman DG**. *Practical statistics for medical research*. New York: Chapman & Hall/CRC, 1991.
7 **Goldberg MS**, Siemiatycki J, Gerin M. Inter-rater agreement in assessing occupational exposure in a case-control study. *Br J Ind Med* 1986;**43**:667–76.
8 **Benke G**, Sim M, Forbes A, *et al.* Retrospective assessment of occupational exposure to chemicals in community-based studies: validity and repeatability of industrial hygiene panel ratings. *Int J Epidemiol* 1997;**26**:635–42.
9 **Siemiatycki J**, Fritschi L, Nadon L, *et al.* Reliability of an expert rating procedure for retrospective assessment of occupational exposures in community-based case–control studies. *Am J Ind Med* 1997;**31**:280–6.
10 **Rybicki BA**, Peterson EL, Johnson CC, *et al.* Intra- and inter-rater agreement in the assessment of occupational exposure to metals. *Int J Epidemiol* 1998;**27**:269–73.
11 **van Tongeren M**, Nieuwenhuijsen MJ, Gardiner K, *et al.* A job-exposure matrix for potential endocrine-disrupting chemicals developed for a study into the association between maternal occupational exposure and hypospadias. *Ann Occup Hyg* 2002;**46**:465–77.
12 **Shrout PE**, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86**:420–8.
13 **McGraw KO**, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996;**1**:30–46.
14 **Macaluso M**, Delzell E, Rose V, *et al.* Inter-rater agreement in the assessment of solvent exposure at a car assembly plant. *Am Ind Hyg Assoc J* 1993;**54**:351–9.
15 **de Cock J**, Kromhout J, Heederik D, *et al.* Experts' subjective assessment of pesticide exposure in fruit growing. *Scand J Work Environ Health* 1996;**22**:425–32.
16 **Stewart PA**, Carel R, Schairer C, *et al.* Comparison of industrial hygienists' exposure evaluations for an epidemiological study. *Scand J Work Environ Health* 2000;**26**:44–51.
17 **Takahashi K**, Case BW, Dufresne A, *et al.* Relation between lung asbestos fibre burden and exposure indices based on job history. *Occup Environ Med* 1994;**51**:461–9.
18 **Segnan N**, Ponti A, Ronco GF, *et al.* Comparison of methods for assessing the probability of exposure in metal plating, shoe and leather goods manufacture and vine growing. *Occup Hyg* 1996;**3**:199–208.
19 **Post W**, Kromhout H, Heederik D, *et al.* Semiquantitative estimates of exposure to methylene chloride and styrene: The influence of quantitative exposure data. *Appl Occup Environ Hyg* 1991;**6**:197–204.
20 **Strand LÅ**, Andersen A. Kartlegging av kreftrisiko og årsaksspesifikk dødelighet blant ansatte i norsk offshorevirksomhet. Innsamling av bakgrunnsdata og etablering av kohort [Survey of cancer risk and cause-specific mortality of Norwegian offshore oil industry workers. Collection of background data and establishment of a cohort]. Oslo: Cancer Registry of Norway, 2001.

21 **Steinsvåg K**, Bråtveit M, Moen BE. Exposure to oil mist and oil vapour during offshore drilling in Norway, 1979–2004. *Ann Occup Hyg* 2006;**50**:109–22.

22 **Steinsvåg K**, Bråtveit M, Moen BE. Exposure to carcinogens for defined job categories in Norway's offshore petroleum industry, 1970–2005. *Occup Environ Med*. 2006; doi: 10, 1136/oem.2006.028225..

23 **Fleiss JL**. *Statistical methods for rates and proportions*. New York: Wiley, 1981:212–36.

24 **Teschke K**, Marion SA, Ostry A, *et al*. Reliability of retrospective chlorophenol exposure estimates over five decades. *Am J Ind Med* 1996;**30**:616–22.

25 **Kromhout H**, Symanski E, Rappaport SM. A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. *Ann Occup Hyg* 1993;**37**:253–70.

26 **Roberts C**, McNamee R. Assessing the reliability of ordered categorical scales using kappa-type statistics. *Stat Methods Med Res* 2005;**14**:493–514.

27 **Fleiss JF**, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas* 1973;**33**:613–19.

APPENDIX

Intraclass correlation coefficients (ICCs), case 2

ICCs, case 2,[12] are calculated using two-way random analysis of variance including a random effect for the score of each set of ratings per combination and a random rater effect for each of the raters.

The two-way random effects variance model is as follows:[12 13]

$x_{ij} = \mu + a_i + b_j + e_{ij}$

where x_{ij} = the ith rating (i = 1,…,k) on the jth combination of agent/job category/time period (j = 1,…,n); μ = the overall population mean of the ratings; a_i = the difference from μ of the mean of the ith rater's ratings; b_j = the difference from μ of the jth combination's so-called true score (that is, the mean across many repeated ratings on the jth combination); and e_{ij} = the random error in the ith rater's scoring of the jth combination. The model assumes that a_i, b_j and e_{ij} are random, independent and normally distributed with mean 0 and variance $\sigma_a^2/\sigma_b^2/\sigma_e^2$.

Table A1 gives the mean square expectations for analysis of the variance model.

Table A2 gives the formulas for single and average score ICC for the two-way variance models. The larger the between-combination variance is with respect to the between-rater variance and residual variance, the higher the ICC. An ICC of 1 indicates perfect inter-rater agreement.

Table A1 Mean square expectations for analysis of the variance model[25]

Source of variation	Degrees of freedom	Mean square	Expected mean square
Between combinations	n–1	$MS_{combination}$	$\sigma_{residual+}k\sigma_{combination}$
Between raters	k–1	MS_{rater}	$\sigma_{residual+}n\sigma_{rater}$
Residual	(n–1)(k–1)	$MS_{residual}$	$\sigma_{residual}$

n, number of combinations; k, number of raters.

Table A2 Single and average score intraclass correlation coefficients (ICCs) for the two-way variance model[13]

Score	Designation	p Value	Formula for calculating	Interpretation of ICC
Single	$ICC_{(2,1)}$	$\frac{\sigma^2_{combination}}{\sigma^2_{combination}+\sigma^2_{rater}+\sigma^2_{residual}}$	$\frac{MS_{combination}-MS_{residual}}{MS_{combination}+(k-1)MS_{residual}+\frac{k}{n}(MS_{rater}-MS_{residual})}$	The degree of absolute agreement between exposure assessments
Average	$ICC_{(2,k)}$	$\frac{\sigma^2_{combination}}{\sigma^2_{combination}+(\sigma^2_{rater}+\sigma^2_{residual})/k}$	$\frac{MS_{combination}-MS_{residual}}{MS_{combination}+(MS_{rater}-MS_{residual})/n}$	The degree of absolute agreement for exposure assessments that are averages based on k independent exposure assessments on randomly selected combinations of exposure